

Capturing, Ordering and Gaussianity in 2D

STEVEN FINCH

January 19, 2016

ABSTRACT. We collect various facts related loosely to random Gaussian quadrilaterals in the plane. For example, a side of a degenerate quadrilateral (one point inside three others) has a density that is non-Rayleigh.

Let A, B, C be independent random Gaussian points in \mathbb{R}^2 , all of which have mean vector zero and covariance matrix identity. The triangle ABC is said to **capture** the origin if $(0, 0)$ is contained in the convex hull of A, B, C . This event occurs with probability $1/4$ and a highly attractive proof appears in [1]. We offer an *unattractive* proof of the same, with the advantage that $(0, 0)$ can be replaced by an arbitrary location (ξ, η) , but the results are available only numerically. As far as is known, a closed-form expression for probabilities does not exist.

We then review the variance of the median of three points in \mathbb{R}^1 and examine how this might be generalized to four points in \mathbb{R}^2 . The convex-hull peeling characterization for **order statistics** (better: **data depth**) in the plane is usually said to be distributionally intractable [2, 3, 4, 5]. The reason will become rather clear here! Multivariate integrals are written down, but no attempt is made to evaluate them.

Finally, we summarize what is known about random Gaussian quadrilaterals in \mathbb{R}^2 and give some simulation-based outcomes. Our hope (as always) is that someone else might be able to break the theoretical logjam and provide a rigorous analysis supporting these.

1. LOCATION CAPTURE

Let $A = (a_1, a_2)$, $B = (b_1, b_2)$, $C = (c_1, c_2)$. The triangle ABC captures the location (ξ, η) if and only if the three determinants [6, 7, 8]

$$\begin{vmatrix} \xi & \eta & 1 \\ b_1 & b_2 & 1 \\ c_1 & c_2 & 1 \end{vmatrix}, \quad \begin{vmatrix} a_1 & a_2 & 1 \\ \xi & \eta & 1 \\ c_1 & c_2 & 1 \end{vmatrix}, \quad \begin{vmatrix} a_1 & a_2 & 1 \\ b_1 & b_2 & 1 \\ \xi & \eta & 1 \end{vmatrix}$$

are all positive or all negative. By circular symmetry of the bivariate normal distribution, we set $\eta = 0$ without loss of generality (upon rotation). Rather than

⁰Copyright © 2016 by Steven R. Finch. All rights reserved.

work with $N(0, 1)$ variables and $(\xi, 0)$, we choose $N(-\xi, 1)$ variables and $(0, 0)$ for simplicity's sake (upon translation). Let us focus on the scenario

$$b_1 c_2 - b_2 c_1 > 0, \quad a_1 b_2 - a_2 b_1 > 0, \quad a_2 c_1 - a_1 c_2 > 0$$

remembering to multiply by 2 later. Within this, there are two cases:

$$a_1 > 0, \quad b_1 > 0, \quad c_1 < 0;$$

$$a_1 < 0, \quad b_1 < 0, \quad c_1 > 0$$

and we must remember to multiply by 3 later.

The first case gives rise to

$$c_2 > \frac{b_2 c_1}{b_1}, \quad b_2 > \frac{a_2 b_1}{a_1}, \quad c_2 < \frac{a_2 c_1}{a_1}$$

and inner integrals

$$\begin{aligned} & \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{\infty} \int_{a_2 b_1 / a_1}^{\infty} \int_{b_2 c_1 / b_1}^{a_2 c_1 / a_1} \exp\left(-\frac{a_2^2 + b_2^2 + c_2^2}{2}\right) dc_2 db_2 da_2 \\ &= \frac{1}{4\pi} \int_{-\infty}^{\infty} \int_{a_2 b_1 / a_1}^{\infty} \exp\left(-\frac{a_2^2 + b_2^2}{2}\right) \left[\operatorname{erf}\left(\frac{a_2 c_1}{\sqrt{2} a_1}\right) - \operatorname{erf}\left(\frac{b_2 c_1}{\sqrt{2} b_1}\right) \right] db_2 da_2 \\ &= \frac{1}{4\sqrt{2}\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{a_2^2}{2}\right) \left[1 + \operatorname{erf}\left(\frac{a_2 c_1}{\sqrt{2} a_1}\right) + 4T\left(\frac{a_2 c_1}{a_1}, \frac{b_1}{c_1}\right) \right] da_2 \\ &= \frac{1}{4} + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-t^2) T(tz, w) dt \end{aligned}$$

where $t = a_2 / \sqrt{2}$, $z = \sqrt{2} c_1 / a_1$, $w = b_1 / c_1$ and $T(h, k)$ is Owen's T function [9, 10]

$$T(h, k) = \frac{1}{2\pi} \int_0^k \frac{1}{1+s^2} \exp\left(-\frac{h^2(1+s^2)}{2}\right) ds.$$

Integrating by parts:

$$\begin{aligned} u &= T(tz, w), & dv &= \frac{1}{\sqrt{\pi}} \exp(-t^2) dt, \\ du &= -\frac{z}{2\sqrt{2}\pi} \exp\left(-\frac{t^2 z^2}{2}\right) \operatorname{erf}\left(\frac{tzw}{\sqrt{2}}\right) dt, & v &= \frac{1}{2} \operatorname{erf}(t) \end{aligned}$$

we obtain

$$\begin{aligned} & \frac{1}{4} + \frac{z}{4\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{t^2 z^2}{2}\right) \operatorname{erf}(t) \operatorname{erf}\left(\frac{t z w}{\sqrt{2}}\right) dt \\ &= \frac{1}{4} + \frac{p}{4\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-p^2 t^2) \operatorname{erf}(t) \operatorname{erf}(q t) dt \end{aligned}$$

where $p = z/\sqrt{2} = c_1/a_1$ and $q = z w/\sqrt{2} = b_1/a_1$. No antiderivative is possible, but the definite integral is known [11], yielding

$$\frac{1}{4} + \frac{p}{4\sqrt{\pi}} \frac{2}{\sqrt{\pi} p} \arctan\left(\frac{q}{p\sqrt{1+p^2+q^2}}\right) = \frac{1}{4} + \frac{1}{2\pi} \arctan\left(\frac{a_1 b_1}{c_1 \sqrt{a_1^2 + b_1^2 + c_1^2}}\right).$$

The second case gives rise to

$$c_2 < \frac{b_2 c_1}{b_1}, \quad b_2 < \frac{a_2 b_1}{a_1}, \quad c_2 > \frac{a_2 c_1}{a_1}$$

and inner integrals

$$\begin{aligned} & \frac{1}{(2\pi)^{3/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{a_2 b_1/a_1} \int_{a_2 c_1/a_1}^{b_2 c_1/b_1} \exp\left(-\frac{a_2^2 + b_2^2 + c_2^2}{2}\right) dc_2 db_2 da_2 \\ &= \frac{1}{4\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{a_2 b_1/a_1} \exp\left(-\frac{a_2^2 + b_2^2}{2}\right) \left[\operatorname{erf}\left(\frac{b_2 c_1}{\sqrt{2} b_1}\right) - \operatorname{erf}\left(\frac{a_2 c_1}{\sqrt{2} a_1}\right) \right] db_2 da_2 \\ &= \frac{1}{4\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{a_2^2}{2}\right) \left[1 - \operatorname{erf}\left(\frac{a_2 c_1}{\sqrt{2} a_1}\right) + 4T\left(\frac{a_2 c_1}{a_1}, \frac{b_1}{c_1}\right) \right] da_2. \end{aligned}$$

The only difference with before is the sign in front of the error function; the integral ultimately simplifies to

$$\frac{1}{4} - \frac{1}{2\pi} \arctan\left(\frac{a_1 b_1}{c_1 \sqrt{a_1^2 + b_1^2 + c_1^2}}\right)$$

because $|a_1| = -a_1$ here.

Multiplying both expressions by 6, the desired probability is

$$\frac{3}{(2\pi)^{5/2}} [\varphi(\xi) + \psi(\xi)] = \begin{cases} 0.250000... & \text{if } \xi = 0, \\ 0.197171... & \text{if } \xi = 1/2, \\ 0.098289... & \text{if } \xi = 1, \\ 0.032455... & \text{if } \xi = 3/2, \\ 0.007626... & \text{if } \xi = 2 \end{cases}$$

where

$$\begin{aligned} \varphi &= \int_0^\infty \int_0^\infty \int_{-\infty}^0 \exp\left(-\frac{(a_1+\xi)^2+(b_1+\xi)^2+(c_1+\xi)^2}{2}\right) \left[\pi + 2 \arctan\left(\frac{a_1 b_1}{c_1 \sqrt{a_1^2+b_1^2+c_1^2}}\right)\right] dc_1 db_1 da_1, \\ \psi &= \int_{-\infty}^0 \int_{-\infty}^0 \int_0^\infty \exp\left(-\frac{(a_1+\xi)^2+(b_1+\xi)^2+(c_1+\xi)^2}{2}\right) \left[\pi - 2 \arctan\left(\frac{a_1 b_1}{c_1 \sqrt{a_1^2+b_1^2+c_1^2}}\right)\right] dc_1 db_1 da_1. \end{aligned}$$

Further symbolic integration does not appear to be possible.

In the preceding analysis, the location (ξ, η) was assumed to be fixed. Suppose instead that X is a random Gaussian point in \mathbb{R}^2 – independent of A, B, C – with mean vector zero and covariance identity. The probability that X falls in the convex hull of A, B, C is [12, 13]

$$-\frac{1}{2} + \frac{3}{2} \frac{\operatorname{arcsec}(3)}{\pi} = 0.0877398280459109052562833... = \frac{1-\theta}{4},$$

the **expected probability content** of ABC . As the name suggests, this is a mean (over all Gaussian triangles ABC in the plane). The corresponding variance is unknown, although recent progress has been made [14, 15]. We will see θ again later.

Similarly, a tetrahedron in \mathbb{R}^3 captures the origin iff four 3×3 determinants are all positive or negative. One sample determinant is

$$\begin{vmatrix} a_1 & a_2 & a_3 & 1 \\ b_1 & b_2 & b_3 & 1 \\ c_1 & c_2 & c_3 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix} = a_1 b_2 c_3 + a_2 b_3 c_1 + a_3 b_1 c_2 - a_1 b_3 c_2 - a_2 b_1 c_3 - a_3 b_2 c_1$$

and the algebraic challenge of working with four such inequalities is clear. The expected probability content of a Gaussian tetrahedron in space is, however, known to be [12, 13]

$$-\frac{2}{5} + \frac{\operatorname{arcsec}(4)}{\pi} = 0.0195693767448337562290498....$$

This quantity is also called the **Gaussian volume** since it is a distribution-dependent analog of ordinary Euclidean measure.

2. MEDIAN VARIANCE

In one dimension, points a and b capture x iff the two determinants

$$\begin{vmatrix} x & 1 \\ b & 1 \end{vmatrix}, \quad \begin{vmatrix} a & 1 \\ x & 1 \end{vmatrix}$$

are both positive or both negative. Let us focus on the scenario $x - b < 0$ and $a - x < 0$, that is, over the domain

$$\Omega_{a,b} = \{x \in \mathbb{R} : a < x < b\}.$$

The desired variance is [16]

$$\frac{6}{(2\pi)^{3/2}} \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\Omega_{a,b}} x^2 \exp\left(-\frac{a^2 + b^2 + x^2}{2}\right) dx db da = 1 - \frac{\sqrt{3}}{\pi}.$$

A simulation suggests that the median is normally distributed; this is *false*, but the true density

$$\frac{6}{(2\pi)^{3/2}} \int_{-\infty}^x \int_x^{\infty} \exp\left(-\frac{a^2 + b^2 + x^2}{2}\right) db da = \frac{3}{2\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \left[1 - \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)^2\right]$$

is nevertheless very close to that for $N(0, 1 - \sqrt{3}/\pi)$.

We have already stated requirements in two dimensions for points A, B, C to capture X . Define

$$\Omega_{A,B,C} = \left\{ \begin{array}{l} X \in \mathbb{R}^2 : (b_1 - a_1)(x_2 - a_2) - (b_2 - a_2)(x_1 - a_1) > 0, \\ (c_1 - b_1)(x_2 - b_2) - (c_2 - b_2)(x_1 - b_1) > 0 \\ \text{and } (a_1 - c_1)(x_2 - c_2) - (a_2 - c_2)(x_1 - c_1) > 0 \end{array} \right\}$$

and let $|X|^2 = x_1^2 + x_2^2$ for convenience. The desired variance is

$$\sigma^2 = \frac{8}{(2\pi)^4(1 - \theta)} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \int_{\Omega_{A,B,C}} x_1^2 \exp\left(-\frac{|A|^2 + |B|^2 + |C|^2 + |X|^2}{2}\right) dX dC dB dA.$$

Simulation suggests normality again holds (approximately) and that $\sigma^2 \approx 0.36$. Evaluating this eight-fold integral, or even removing just x_2 , does not seem feasible. The factor involving

$$1 - \theta = -2 + 6 \frac{\operatorname{arcsec}(3)}{\pi} = 0.3509593121836436210251333...$$

is necessary for reasons to be made clear shortly.

3. GAUSSIAN QUADRILATERALS

Let A, B, C, D be independent random Gaussian points in \mathbb{R}^2 , all of which have mean vector zero and covariance matrix identity. The convex hull $ABCD$ of A, B, C, D possesses either three or four vertices; if the former is true, $ABCD$ is a **degenerate quadrilateral**; if the latter is true, $ABCD$ is a **quadrilateral**. (Non-degeneracy rules out triangles.) The probability that the number of vertices is four is [12, 13]

$$\theta = 3 - 6 \frac{\text{arcsec}(3)}{\pi} = 0.6490406878163563789748666....$$

Earlier occurrences of θ can now be explained. In Section 1, $(1-\theta)/4$ is the probability that a specific point (say, D) is inside the other three (A, B, C). By contrast, $1-\theta$ is the probability that *any* one of the four points is internal to the rest. In Section 2, given three points on a line, a median always exists (thus the denominator is 1). By contrast, given four points in the plane, we have implicitly conditioned on degeneracy so that an inner point exists (thus the denominator here is $1-\theta$).

With no constraints, the convex hull $ABCD$ has expected area $\sqrt{3}$ and expected perimeter

$$(3+\theta)\sqrt{\pi} = 6.4677562192310137839669010....$$

It is interesting that $\sqrt{3}$ is twice the expected area of a planar Gaussian triangle [17, 18, 19, 20], but the corresponding triangular perimeter $3\sqrt{\pi}$ is not so simply related to $(3+\theta)\sqrt{\pi}$. No higher moments for either area of $ABCD$ or perimeter of $ABCD$ are precisely known.

If the number of vertices of $ABCD$ is constrained to be three, then

$$E(\text{side}) \approx 2.11, \quad E(\text{side}^2) \approx 5.32.$$

(By “side” is meant the distance between two adjacent vertices on the exterior.) We deduce that the side density here is non-Rayleigh as $E(\text{side})^2 / E(\text{side}^2)$ is not sufficiently near to $\pi/4$. If the number of vertices of $ABCD$ is constrained to be four, then

$$E(\text{side}) \approx 1.45, \quad E(\text{side}^2) \approx 2.78.$$

We are also interested in cross-correlations between sides. For the case of quadrilaterals, how does the cross-correlation between sides sharing a vertex differ from the cross-correlation between sides that are wholly disjoint?

Problems involving random quadrilaterals have occupied our attention for many years [21] and remain essentially unsolved. Much more relevant material can be found at [22], including experimental computer runs that aided theoretical discussion here.

REFERENCES

- [1] R. Howard and P. Sisson, Capturing the origin with random points: generalizations of a Putnam problem, *College Math. J.* 27 (1996) 186–192; MR1390366.
- [2] M. I. Shamos, Geometry and statistics: problems at the interface, *Algorithms and Complexity*, Proc. Carnegie-Mellon Univ. sympos., ed. J. F. Traub, Academic Press, 1976, pp. 251–280; MR0431785 (55 #4780).
- [3] V. Barnett, The ordering of multivariate data, *J. Royal Statist. Soc. Ser. A* 139 (1976) 318–355; MR0445726 (56 #4060).
- [4] G. Aloupis, Geometric measures of data depth, *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, Proc. 2003 Rutgers Univ. workshop, ed. R. Y. Liu, R. Serfling and D. L. Souvaine, Amer. Math. Soc., 2006, pp. 147–158; MR2343118.
- [5] J. Hugg, E. Rafalin, K. Seyboth and D. Souvaine, An experimental study of old and new depth measures, *Proc. 2006 Workshop on Algorithm Engineering and Experiments (ALENEX)*, Miami, ed. R. Raman and M. F. Stallmann, SIAM, pp. 51–64.
- [6] G. Herron, Point within a tetrahedron, USENET comp.graphics.algorithms posting, 1994.
- [7] Anonymous contributor, Determining if an arbitrary point lies inside a triangle defined by three points, Mathematics Stack Exchange posting, 2011.
- [8] Anonymous contributor, How to check whether the point is in the tetrahedron or not, Stack Overflow posting, 2014.
- [9] D. B. Owen, Tables for computing bivariate normal probabilities, *Annals Math. Statist.* 27 (1956) 1075–1090; MR0127562 (23 #B607).
- [10] H. A. Fayed and A. F. Atiya, An evaluation of the integral of the product of the error function and the normal probability density with application to the bivariate normal integral, *Math. Comp.* 83 (2014) 235–250; MR3120588.
- [11] A. Dieckmann, Table of Integrals, <http://www-elsa.physik.uni-bonn.de/~dieckman/IntegralsDefinition>
- [12] B. Efron, The convex hull of a random set of points, *Biometrika* 52 (1965) 331–343; MR0207004 (34 #6820).

- [13] S. N. Majumdar, A. Comtet and J. Randon-Furling, Random convex hulls and extreme value statistics, *J. Stat. Phys.* 138 (2010) 955–1009; MR2601420 (2011c:62166).
- [14] C. Buchta, An identity relating moments of functionals of convex hulls, *Discrete Comput. Geom.* 33 (2005) 125–142; MR2105754 (2005j:60022).
- [15] M. Beermann and M. Reitzner, Beyond the Efron-Buchta identities: distributional results for Poisson polytopes, *Discrete Comput. Geom.* 53 (2015) 226–244; MR3293497.
- [16] H. L. Jones, Exact lower moments of order statistics in small samples from a normal distribution, *Annals Math. Statist.* 19 (1948) 270–273; MR0025121 (9,601d).
- [17] S. R. Finch, Random triangles, unpublished note (2010), <http://www.people.fas.harvard.edu/~sfinch/>.
- [18] P. Clifford and N. J. B. Green, Distances in Gaussian point sets, *Math. Proc. Cambridge Philos. Soc.* 97 (1985) 515–524; MR0778687 (86i:62091).
- [19] K. S. Miller, Complex Gaussian processes, *SIAM Rev.* 11 (1969) 544–567; MR0258109 (41 #2756).
- [20] K. S. Miller, *Complex Stochastic Processes. An Introduction to Theory and Application*, Addison-Wesley, 1974, pp. 86–100; MR0368118 (51 #4360).
- [21] W. J. C. Miller, Problem 6543, *The Educational Times* 33 (1880) 311; 53 (1900) 225.
- [22] S. R. Finch, Simulations in R involving quadrilaterals, <http://www.people.fas.harvard.edu/~sfinch/csolve/rsimul.html>.

Steven Finch
Dept. of Statistics
Harvard University
Cambridge, MA, USA
steven_finch@harvard.edu